

# A matter of words: NLP for quality evaluation of Wikipedia medical articles

An extended abstract of this work will appear as a short paper in  
proceedings of ICWE2016, published by LNCS Springer.

Vittoria Cozza<sup>1</sup>, Marinella Petrocchi<sup>1</sup>, and Angelo Spognardi<sup>2</sup>

<sup>1</sup> IIT CNR, Pisa, Italy {v.cozza, m.petrocchi}@iit.cnr.it

<sup>2</sup> DTU Lingby, Denmark angsp@dtu.dk

**Abstract.** Automatic quality evaluation of Web information is a task with many fields of applications and of great relevance, especially in critical domains like the medical one. We move from the intuition that the quality of content of medical Web documents is affected by features related with the specific domain. First, the usage of a specific vocabulary (Domain Informativeness); then, the adoption of specific codes (like those used in the infoboxes of Wikipedia articles) and the type of document (e.g., historical and technical ones). In this paper, we propose to leverage specific domain features to improve the results of the evaluation of Wikipedia medical articles. In particular, we evaluate the articles adopting an “actionable” model, whose features are related to the content of the articles, so that the model can also directly suggest strategies for improving a given article quality. We rely on Natural Language Processing (NLP) and dictionaries-based techniques in order to extract the biomedical concepts in a text. We prove the effectiveness of our approach by classifying the medical articles of the Wikipedia Medicine Portal, which have been previously manually labeled by the Wiki Project team. The results of our experiments confirm that, by considering domain-oriented features, it is possible to obtain sensible improvements with respect to existing solutions, mainly for those articles that other approaches have less correctly classified. Other than being interesting by their own, the results call for further research in the area of domain specific features suitable for Web data quality assessment.

## 1 Introduction

As observed by a recent article of Nature News [16], “Wikipedia is among the most frequently visited websites in the world and one of the most popular places to tap into the world’s scientific and medical information”. Despite the huge amount of consultations, open issues still threaten a fully confident fruition of the popular online open encyclopedia.

A first issue relates to the reliability of the information available: since Wikipedia can be edited by anyone, regardless of their level of expertise, this

tends to erode the average reputation of the sources, and, consequently, the trustworthiness of the contents posted by those sources. In an attempt to fix this shortcoming, Wikipedia has recently enlisted the help of scientists to actively support the editing on Wikipedia [16]. Furthermore, lack of control may lead to the publication of fake Wikipedia pages, which distort the information by inserting, e.g., promotional articles and promotional external links. Fighting vandalism is one of the main goals of the Wikimedia Foundation, the nonprofit organization that supports Wikipedia: machine learning techniques have been considered to offer a service to “judge whether an edit was made in good faith or not” [23]. Nonetheless, in the past recent time, malicious organisations have acted disruptively with purposes of extortion - see, e.g., the recent news on the uncovering of a blackmail network of accounts, which threatened celebrities with the menace of inserting offending information on their Wikipedia pages<sup>3</sup>.

Secondly, articles may suffer from readability issues: achieving a syntactical accuracy that helps the reader with a fluid reading experience is —quite obviously— a property which articles should fulfill. Traditionally, the literature has widely adopted well known criteria, as the “Flesch-Kincaid” measure” [17], to automatically assess readability in textual documents. More recently, new techniques have been proposed too, for assessing the readability of natural languages (see, e.g., [13] for the Italian use case, [24] for the Swedish one, [27] for English).

In this paper, we face the quest for quality assessment of a Wikipedia article, in an automatic way that comprehends not only readability and reliability criteria, but also additional parameters testifying completeness of information and coherence with the content one expects from an article dealing with specific topics, plus sufficient insights for the reader to elaborate further on some argument. The notion of data quality we deal with in the paper is coherent with the one suggested by recent contributions (see, e.g., [20]), which points out like the quality of Web information is strictly connected to the scope for which one needs such information.

Our intuition is that groups of articles related to a specific topic and falling within specific scopes are intrinsically different from other groups on different topics within different scopes. We approach the article evaluation through machine learning techniques. Such techniques are not new to be employed for automatic evaluation of articles quality. As an example, the work in [28] exploits classification techniques based on structural and linguistic features of an article. Here, we enrich that model with novel features that are domain-specific. As a running scenario, we focus on the Wikipedia medical portal. Indeed, facing the problems of information quality and ensuring high and correct levels of informativeness is even more demanding when health aspects are involved. Recent statistics report that Internet users are increasingly searching the Web for health information, by consulting search engines, social networks, and specialised health portals, like that of Wikipedia. As pointed out by the 2014 Eurobarometer survey

---

<sup>3</sup> [https://en.wikipedia.org/wiki/Wikipedia:Long-term\\_abuse/Orangemoody](https://en.wikipedia.org/wiki/Wikipedia:Long-term_abuse/Orangemoody)

on European citizens’ digital health literacy<sup>4</sup>, around six out of ten respondents have used the Internet to search for health-related information. This means that, although the trend in digital health literacy is growing, there is also a demand for a qualified source where people can ask and find medical information which, to an extent, can provide the same level of familiarity and guarantees as those given by a doctor or a health professional.

We anticipate here that leveraging new domain-specific features is in line with this demand of articles quality. Moreover, as the outcomes of our experiments show, they effectively improve the classification results in the hard task of multi-class assessment, especially for those classes that other automatic approaches worst classify. Remarkably, our proposal is general enough to be easily extended to other domains, in addition to the medical one.

Section 2 first describes the structure of the articles present in the medical portal. Then, it gives details on the real data used in the experiments, which are indeed articles extracted from the medical portal and labeled according to the manual assessment by the Wikimedia project. Section 3 briefly presents the actionable model in [28]: we adopt it as the baseline for our analysis. In Section 4, we present the domain-specific, medical model we newly adopt in this paper as an extension of the baseline. The extended model includes features specifically extracted from the medical domain. One novel feature is based on the article textual content. Section 5 presents the process which its extraction relies on, with a non trivial analysis of natural language and domain knowledge. Section 6 presents experiments and results, with a comparison of the baseline model with the new one. In Section 7, we survey related work in the area and in Section 8 we conclude the paper.

## 2 Dataset

We consider the dataset consisting of the entire collection of articles of the Wikipedia Medicine Portal, updated at the end of 2014. Wikipedia articles are written according to the Media Wiki markup language, a HTML-like language. Among the structural elements of one page, which differs from standard HTML pages, there are *i)* the internal links, i.e., links to other Wikipedia pages, different from links to external resources); *ii)* categories, which represent the Media Wiki categories a page belongs to: they are encoded in the part of text within the Media Wiki “categories” tag in the page source, and *iii)* informative boxes, so called “infoboxes”, which summarize in a structured manner some peculiar pieces of information related the topic of the article. The category values for the articles in the medical portal span over the ones listed at <https://en.wikipedia.org/wiki/Portal:Medicine>. Examples of categories, which appear at the bottom of each Wikipedia page, are in Fig. 1.

Infoboxes of the medical portal feature medical content and standard coding. As an example, Fig. 2 shows the infobox in the Alzheimer’s disease page of

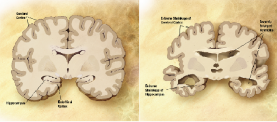
---

<sup>4</sup> [http://ec.europa.eu/public\\_opinion/flash/fl\\_404\\_sum\\_en.pdf](http://ec.europa.eu/public_opinion/flash/fl_404_sum_en.pdf)

Categories: [Alzheimer's disease](#) | [Ailments of unknown etiology](#) | [Unsolved problems in neuroscience](#)  
[Learning disabilities](#) | [Psychiatric diagnosis](#) | [Dementia](#) | [Abnormal psychology](#) | [Cognitive disorders](#)  
[Aphasias](#) | [Herpes simplex virus-associated diseases](#)

**Fig. 1.** Example of Wikipedia Medicine Portal article categories

the portal. The infobox contains explanatory figures and text denoting peculiar characteristics of the disease and the value for the standard code of such disease (ICD9, as for the international classification of the disease<sup>5</sup>).

Alzheimer's disease	
 <p>Comparison of a normal aged brain (left) and the brain of a person with Alzheimer's (right). Characteristics that separate the two are pointed out.</p>	
Classification and external resources	
<b>Specialty</b>	Neurology
<b>ICD-10</b>	G30 <a href="#">↗</a> , F00 <a href="#">↗</a>
<b>ICD-9-CM</b>	331.0 <a href="#">↗</a> , 290.1 <a href="#">↗</a>
<b>OMIM</b>	104300 <a href="#">↗</a>
<b>DiseasesDB</b>	490 <a href="#">↗</a>
<b>MedlinePlus</b>	000760 <a href="#">↗</a>
<b>eMedicine</b>	neuro/13 <a href="#">↗</a>
<b>Patient UK</b>	Alzheimer's disease <a href="#">↗</a>
<b>MeSH</b>	D000544 <a href="#">↗</a>
<b>GeneReviews</b>	NBK1161 <a href="#">↗</a>

**Fig. 2.** The infobox on Alzheimer's disease

Thanks to WikiProject Medicine<sup>6</sup>, the dataset of articles we collected from the Wikipedia Medicine Portal has been manually labeled into seven quality classes. They are ordered as *Stub*, *Start*, *C*, *B*, *A*, *Good Article (GA)*, *Featured Article (FA)*. The Featured and Good article classes are the highest ones: to have those labels, an article requires a community consensus and an official review by selected editors, while the other labels can be achieved with reviews from a larger, even controlled, set of editors. Actually, none of the articles in the dataset is labeled as *A*, thus, in the following, we do not consider that class, restricting the investigation to six classes.

<sup>5</sup> <http://www.who.int/classifications/icd/en/>

<sup>6</sup> [https://en.wikipedia.org/wiki/Wikipedia:WikiProject\\_Medicine/Assessment](https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Assessment)

At the date of our study, we were able to gather 24,362 rated documents. Remarkably, only a small percentage of them (1%) is labeled as *GA* and *FA*. Indeed, the distribution of the articles among the classes is highly skewed. There are very few (201) articles for the highest quality classes (*FA* and *GA*), while the vast majority (19,108) belongs to the lowest quality ones (*Stub* and *Start*). This holds not only for the medical portal. Indeed, it is common in all Wikipedia, where, on average, only one article in every thousand is a Featured one.

In Section 6, we will adopt a set of machine-learning classifiers to automatically label the articles into the quality classes. Dealing with imbalanced classes is a common situation in many real applications of classification learning: healthy patients over the population, fraudulent actions over daily genuine transactions, and so on. Without any countermeasure, common classifiers tend to correctly identify only articles belonging to the majority classes, clearly leading to severe mis-classification of the minority classes, since typical learning algorithms strive to maximize the overall prediction accuracy. To reduce the disequilibrium among the size of the classes, we have first randomly sampled the articles belonging to the most populated classes. Then, we have performed some further elaboration, as shown in the following.

Many studies have been conducted to improve learning algorithms accuracy in presence of imbalanced data [15]. For the current work, we have considered one of the most popular approaches, namely the Synthetic Sampling with Data Generation, detailed in [9]. It consists in generating synthetic instances from the minority classes, to balance the overall dataset. The approach has been broadly applied to problems relying on NLP features, see, e.g., [10]. In our case, we re-sampled the input data set by applying the Synthetic Minority Oversampling TEchnique (SMOTE<sup>7</sup>), with percentage 40% for *GA* and 180%, for *FA*. In particular, the steps to oversample are the following:

- New instances are generated using as seed real examples from the minority class;
- For each real example, its  $k$  ( $k = 5$ ) nearest neighbours examples are identified;
- Synthetic instances are generated to be at a random point between the seed and the neighbours.

Table 1 shows the number of articles in the dataset, divided per class, as well as the random samples we have considered for our study. The experiments presented in Section 6 are based on the articles of the right-hand column in the table.

### 3 Baseline: the actionable model

We apply a multi-class classification approach to label the articles of the sampled dataset into the six WikiProject quality classes. In order to have a baseline, we first apply the state of the art model proposed in [28] to the dataset.

---

<sup>7</sup> Implemented and available in the Weka framework

class	original dataset	with majority classes sampling	with minority classes oversampling
Stub	9,267	1,015	1,015
Start	9,841	1,015	1,015
C	3,149	1,015	1,015
B	1,894	1,015	1,015
GA	153	153	214
FA	58	58	162
total	24,362	4,271	4,436

**Table 1.** Dataset

The “actionable model” in [28] focuses on five linguistic and structural features and it weighs them as follows:

1.  $\text{Completeness} = 0.4 * \text{NumBrokenWikilinks} + 0.4 * \text{NumWikilinks}$
2.  $\text{Informativeness} = 0.6 * \text{InfoNoise} + 0.3 * \text{NumImages}$
3.  $\text{NumHeadings}$
4.  $\text{ArticleLength}$
5.  $\text{NumReferences} / \text{ArticleLength}$

where

- $\text{NumWikilinks}$  is the number of links pointing to other Wikipedia pages (whereas  $\text{NumBrokenWikilinks}$  counts the links that are broken);
- $\text{InfoNoise}$  is the proportion of text content remaining in the article after removing MediaWiki markups and cleaning the text with basic NLP operations, such as stopwords removal;
- $\text{ArticleLength}$  is the base 10 log of the article length in bytes;
- $\text{NumHeadings}$ ,  $\text{NumReferences}$  and  $\text{NumImages}$  are, quite intuitively, the number of headers, references and images that an article contains.

In order to evaluate such model over our dataset, we have extracted from our the dataset the above mentioned features. We have measured  $\text{ArticleLength}$ ,  $\text{NumWikilinks}$  and  $\text{NumBrokenWikilinks}$  as suggested in [12].

As shown in Figure 3, the actionable model features have been extracted applying simple scripts mainly based on regular expressions (the “regex extractor” block), which process the whole HTML code of an article (free text plus media wiki tags). The regex extractor relies on Python *BeautifulSoup*<sup>8</sup> for extracting HTML structures and excerpts of the textual content within the MediaWiki tags and on *nlTK* libraries<sup>9</sup> for basic NLP analysis. In details, *nlTK* has been used for computing the InfoNoise feature, whose computation includes the stopwords removal, following the Porter Stopwords Corpus available through *nlTK* [5].

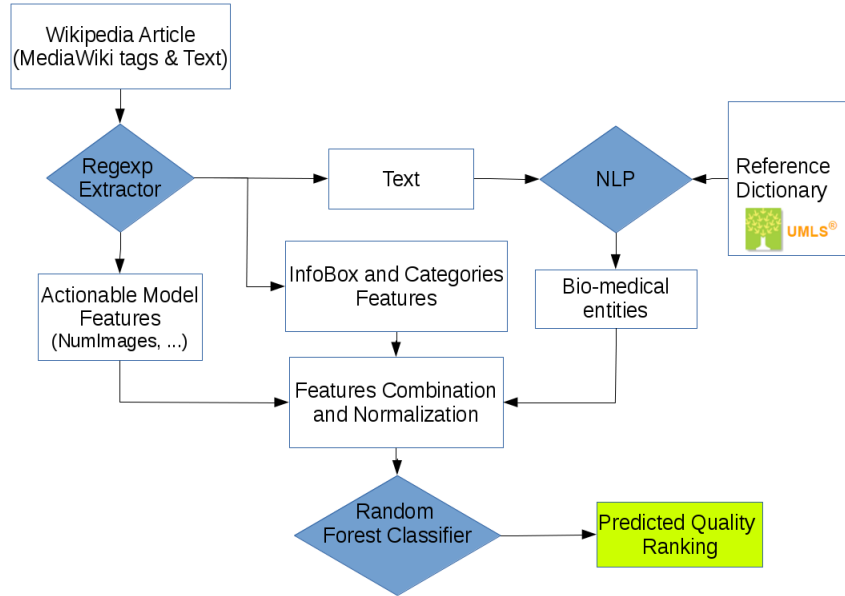
<sup>8</sup> <http://www.crummy.com/software/BeautifulSoup/>

<sup>9</sup> <http://www.nltk.org/>

The classification results according to the baseline model are reported in Section 6.

## 4 The medical domain model

Here, we improve the baseline model with novel and specifically crafted features that rely on the medical domain and that capture details on the specific content of an article. As shown in Figure 3, medical model features, the bio-medical entities, have been extracted from the free text only, exploiting advanced NLP techniques and using domain dictionaries.



**Fig. 3.** Quality Assessment

In details, we newly define and extract from the dataset the following novel features:

1. *InfoBoxNormSize*: this feature represents the normalised size of an infobox that contains standard medical coding.
2. *Category*: the category a page belongs to.
3. *DomainInformativeness*: the number of bio-medical entities, which are the domain dependent terms in the article (such as the ones denoting symptoms, diseases, treatments, etc.).

The idea of considering infoboxes is not novel: for example, in [28] the authors noticed that the presence of an infobox is a characteristic featured by good articles. However, in the specific case of the Medicine Portal, the presence of an infobox does not seem strictly related to the quality class the article belongs to (according to the manual labelling). Indeed, it is recurrent that articles, spanning all classes, have an infobox, containing a schematic synthesis of the article. In particular, pages with descriptions of diseases usually have an infobox with the medical standard code of the disease (i.e., IDC-9 and IDC-10), as in Figure 2.

As done for the baseline, also the first two features of the medical model have been extracted with ad hoc Python scripts, extracting HTML structures and excerpts of the textual content within the MediaWiki tags.

For their extraction of the bio-medical entities, we consider the textual part of the article only, obtained after removing the MediaWiki tags, and we apply a NLP analysis, which is presented in Section 5.

#### 4.1 Infobox-based feature

We have calculated the Infobox size as the base 10 log of the bytes of data contained within the mediawiki tags that wrap an infobox, and we have normalized it with respect to the ArticleLength, introduced in Section 3.

#### 4.2 Category-based feature

We have leveraged the categories assigned to articles in Wikipedia, in particular relating to the medicine topics available at <https://en.wikipedia.org/wiki/Portal:Medicine>.

category	list of keywords
A	anatom*, embryolog*, organ, tissue
B	born, death, birth
D	disorder, disease, pathology
F	first aid

**Table 2.** Categories

We have defined 4 upper level categories of our interest:

- A *anatomy*: an article is about anatomy;
- B *biography*: an article is a biography of someone or tell the history of something;
- D *disorder*: it is about a disorder;
- F *first aid*: it reports information for first aid or emergency contacts;
- O *other*: none of the above.

We have matched the article’s text within the MediaWiki categories tag with an approximate list of keywords that are related to our category of interest, as reported in Table 2.



## 5 Bio-medical entities

In the literature, there are several methods available for extracting bio-medical entities from a text (i.e., from medical notes and/or articles). We refer to [19] for an overview of valuable existing techniques. In this work, we have adopted a dictionary-based approach, which exploits lexical features and domain knowledge extracted from the Unified Medical Languages System (UMLS) Metathesaurus [7]. The approach has been proposed for the Italian language in a past work [1]. Since the approach combines the usage of linguistic analysis and domain resources, we were able to conveniently adapt it for the English language, being both the linguistic pipeline and UMLS available for multiple languages (including English and Italian).

Dictionary-based approaches have been proved valid for the task of entities' extraction, see, for example, another well known, similar approach to the one adopted here, i.e., Metamap<sup>10</sup>. It is worth noting like, even though dictionary-based approaches could be less precise than Named Entity Recognition [19], in our context even an approximate solution is enough, since we are not annotating medical records. Instead, we are quantifying the mole of inherent information within a text.

### 5.1 Reference dictionary

Several ontologies or taxonomies related to the medical domain are available in English. To build a medical dictionary, we have extracted definitions of medical entities from the Unified Medical Languages System (UMLS) Metathesaurus [7]. UMLS integrates bio-medical resources, such as SNOMED-CT<sup>11</sup> that provides the core terminology for electronic health records. In addition, UMLS also provides a semantic network where each entity in the Metathesaurus has an assigned Concept Unique Identifier (CUI) and it is semantically typed.

From UMLS, we have extracted the entries belonging to the following SNOMED-CT semantic groups: *Treatment*, *Sign or Symptom*, *Disease or Syndrome*, *Body Parts*, *Organs*, or *Organ Components*, *Pathologic Function*, and *Mental or Behavioral Dysfunction*, for a total of more than one million entries, as shown in Table 3 (where the two last semantic groups have been grouped together, under *Disorder*). Furthermore, we have extracted common Drugs and Active Ingredients definitions from RxNorm<sup>12</sup>, accessed by RxTerm<sup>13</sup>.

Starting from the entries in Table 3, we have also computed approximate definitions, exploiting syntactic information of the same entries. In details, we have pre-processed the entries by mean of the *Tanl pipeline* [2], a suite of modules for text analytics and NLP, based on machine learning. Preprocessing has consisted in first dividing the entries into single word forms. Then, for each form, we have

<sup>10</sup> <http://metamap.nlm.nih.gov/>

<sup>11</sup> <http://www.ihtsdo.org/snomed-ct/>

<sup>12</sup> <https://www.nlm.nih.gov/research/umls/rxnorm/>

<sup>13</sup> <https://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTerm.cfm>

semantic groups	definitions
Treatment	671,349
Sign or Symptom	43,779
Body Parts, Organs, or Organ Components	234,075
Disorder	402,298
Drugs	5,109
Active Ingredients	2,774

**Table 3.** Dictionary Composition

identified the lemma (when available) and the part of speech (POS). Thus, we have created an approximate definition that consists in using only the lemma and cleaning the text, excluding punctuation, prepositions and articles. Also, approximate definitions have been normalized by lowercasing each word. As an example, the Disorder entry “aneurysm of the vein of galen” has been stored in the dictionary, along with its approximate definition “aneurysm vein galen”.

## 5.2 Extraction of bio-medical entities

We have extracted the bio-medical entities present in the Wikipedia medical articles through a n-gram-based technique.

A pre-processing phase occurs in a similar way as for the dictionary composition. Given a Wikipedia article written in English, we have pre-processed the textual part through the *Tanl pipeline*. Similar to what described in Section 5.1 for the reference dictionary, we have first divided the text in sentences and the sentences into single word forms. For each form, we have considered the lemma (when available) and the part of speech (POS). For instance, starting from an example sentence extracted from the Wikipedia page on the Alzheimer’s disease: “*Other risk factors include a history of head injuries, depression, or hypertension.*”, we have obtained the annotation shown in Figure 4. As in the case of the dictionary, each word in the text has been lowercasing.

After pre-processing the text of each article, we have attempted to match each n-gram (with n between 1 and 10) in the corpus with the entries in the extended dictionary. We both attempt an exact match and an approximate match, the latter removing prepositions, punctuations and articles from the n-grams. Approximate matching leads to several advantages. Indeed, exploiting the text pre-processing allows to identify dictionary definitions present in the text, even when the number differs. As an example, the dictionary definition “injury” will match with “injuries”, mentioned in the text, because in the approximation one can consider the lemmas. Further, considering the POS allows to identify mentions when interleaved by prepositions, articles, and conjunctions that change the form but do not alter the meaning. As an example, the approximate definition “aneurysm vein galen” will match also with the following n-gram: “the aneurysm and vein of galen”, if present in the text.

Form	Lemma	POS
Other	other	JJ
risk	risk	NN
factors	factor	NNS
include	include	VBP
a	a	DT
history	history	NN
of	of	IN
head	head	NN
injuries	injury	NNS
depression	depression	NN
or	or	CC
hypertension	hypertension	NN

**Fig. 4.** Annotation of a sentence with the Tanl English pipeline

## 6 Experiments and results

In this section, we describe the experiments and report the results for the classification of Wikipedia medical articles into the six classes of the Wikipedia Medicine Portal. We compare the results obtained adopting four different classifiers: the actionable model in [28] and three classifiers that leverage the ad-hoc features from the medical domain discussed in the previous sections. All the experiments were realized within the Weka framework [14] and validated through 10 fold cross-validation.

For each experiment, we relied on the dataset presented in Section 2, and specifically, on that obtained after sampling the majority classes and oversampling the minority ones (right-hand column in Table 1). The dataset serves both as training and test set for the classifiers.

Moreover, to take into account the imbalanced data, we have applied several classification algorithms and, for the sake of conciseness, hereafter we report only the best results we have achieved. In particular, we have experimented with bagging, adaptive boosting and random forest and we report the results for the latter only.

### 6.1 Classifiers’ features

In Table 4, we report a summary of the features for each of the considered models: the baseline model in [28] and two new models that employ the medical domain features. In the *Medical Domain* model, we add to the baseline features the Domain Informativeness, as described in Section 4 and 5. In addition, the *Full Medical Domain* model also considers the features InfoBoxNormSize and Category.

For each of the features, the table also reports the Information Gain, evaluated on the whole dataset (24,362 articles). Information Gain is a well-known metric to evaluate the dependency of one class from a single feature, see, e.g., [11].

Baseline	Medical Domain	Full	Info Gain
		Medical Domain	
ArticleLength	ArticleLength	ArticleLength	0.939
NumHeadings	NumHeadings	NumHeadings	0.732
Completeness	Completeness	Completeness	0.724
NumRef/Length	NumRef/Length	NumRef/Length	0.621
Informativeness	Informativeness	Informativeness	0.377
	DomainInformativ.	DomainInformativ.	0.751
		InfoBoxNormSize	0.187
		Category	0.017

**Table 4.** Classifiers: Features and Information Gain

We can observe how the Domain Informativeness feature has a considerably higher infogain value when compared with Informativeness. We anticipate here that this will lead to a more accurate classification results for the highest classes, as reported in the next section. Leading to a greater accuracy is also true for the other two new features that, despite showing lower values of infogain, are able to further improve the classification results, mainly for the articles belonging to the lowest quality classes (Stub and Start).

## 6.2 Classification results

Table 5 shows the results of our multi-class classification. For each of the classes, we have computed the *ROC Area* and *F-Measure* metrics [21]. The latter, in particular, is usually considered a significant metric in terms of classification, since it combines in one single value all the four indicators that are generally implied for evaluating the classifier performance (i.e., number of True Positives, False Positives, True Negatives and False Negatives). In our scenario, the meaning of the indicators, for each class, are as follows:

- *True Positives* are the articles classified as belonging to a certain class, that indeed belong to that class (according to the quality ratings given by the WikiProject Medicine);
- *True Negatives* are the articles classified as not belonging to a certain class, that indeed do not belong to that class;
- *False Positives* are the articles classified as belonging to a certain class, that do not belong to that class;
- *False Negatives* are the articles classified as not belonging to a certain class, that instead belong to that class.

At a first glance, we observe that, across all the models, the articles with the lowest classification values, for both ROC and F-Measure, are those labeled C and GA. Adding the Domain Informativeness feature produces a classification, which is slightly worse for C and FA articles, but better for the other four classes. This is particularly evident for the F-Measure of the articles of the GA class. A

Metric	Baseline	Medical Domain	Full Medical Domain
ROC Area Stub	0.981	0.982	<b>0.983</b>
ROC Area Start	0.852	0.853	<b>0.858</b>
ROC Area C	0.749	0.747	<b>0.76</b>
ROC Area B	0.825	0.832	<b>0.836</b>
ROC Area GA	0.825	0.908	<b>0.916</b>
ROC Area FA	0.977	0.976	<b>0.978</b>
F-Measure Stub	0.886	<b>0.891</b>	0.89
F-Measure Start	0.587	0.582	<b>0.598</b>
F-Measure C	0.376	0.367	<b>0.397</b>
F-Measure B	0.527	0.541	<b>0.542</b>
F-Measure GA	0.245	0.338	<b>0.398</b>
F-Measure FA	0.634	0.631	<b>0.641</b>

**Table 5.** Classification Results. In bold, the best results.

noticeable major improvement is obtained with the introduction of the features InfoBoxNormSize and Category in the *Medical Domain* model. The ROC Area increases for the articles of all the classes within the *Full Medical Domain*, while the F-Measure is always better than the *Baseline* and almost always better than the *Medical Domain*.

The size of an article, expressed either as the word count, analyzed in [6], or as the article length, as done here, appears a very strong feature, able to discriminate the articles belonging to the highest and lowest quality classes. This is testified also by the results achieved exploiting the baseline model of [28], which poorly succeeds in discriminating the articles of the intermediate quality classes, while achieving good results for Stub and FA. Here, the newly introduced features have a predominant effect on the articles of the highest classes. This could be justified by the fact that those articles contain, on average, more text and, then, NLP-based features can exploit more words belonging to a specific domain.

Then, we observe that the ROC Area and the F-Measure are not tightly coupled (namely: high values for the first metric can correspond to low values for the second one, see for example C and GA): this is due to the nature of the ROC Area, that is affected by the different sizes of the considered classes. As an example, we can observe that the baseline model has the same ROC Area value for the articles of both class B and class GA, while the F-Measure of articles of class B is 0.282 higher than that of class GA.

Finally, the results confirm that the adoption of domain-based features and, in general, of features that leverage NLP, help to distinguish between articles in the lowest classes and articles in the highest classes, as highlighted in bold in Table 5. We notice also that exploiting the full medical domain leads us to the achievement of the best results.

Even if preliminary, we believe that the results are promising and call both for features’ further refinement and novel features, able to discriminate among the intermediate classes too.

## 7 Related work

Automatic quality evaluation of Wikipedia articles has been addressed in previous works with both unsupervised and supervised learning approaches. The common idea of most of the existing work is to identify a feature set, having as a starting point the Wikipedia project guidelines, to be exploited with the objective in mind to distinguish Featured Articles. In [26], Stvilia *et al.* identify a relevant set of features, including lingual, structural, historical and reputational aspects of each article. They show the effectiveness of their metrics by applying both clustering and classification. As a result, more than 90% of FA are correctly identified.

Blumenstock [6] inspects the relevance of the *word-count* feature at each quality stage, showing that it can play a very important role in the quality assessment of Wikipedia articles. Only using this feature, the author achieves a F-measure of 0.902 in the task of classifying featured articles and 0.983 in the task of classifying non featured articles. The best results of the investigation are achieved with a classifier based on a neural network implemented with a multi-layer perceptron.

In [30], the authors try to analyze the factors affecting the quality of Wikipedia articles, with respect to their quality class. The authors evaluate a set of 28 features, over a random sample of 500 Wikipedia articles, by weighing each metric in different stages using neural networks. Findings are that linguistic features weigh more in the lowest quality classes, and structural features, along with historical ones, become more important as the article quality improves. Their results indicate that the information quality is mainly affected by completeness, and to be “well-written” is a basic requirement in the initial stage. Instead, reputation of authors or editors is not so important in Wikipedia because of its horizontal structure. In [29], the authors consider the quality of the data in the infoboxes of Wikipedia, finding a correlation between the quality of information in the infobox and the article itself.

In [28], the authors deal with the problem of discriminating between two large classes, namely *NeedWork*, *GoodEnough* (including in GoodEnough both GA and FA), in order to identify which articles need further revisions for being featured. They also introduce new composite features, those that we have referred to as an “actionable model” in Section 3. They obtain good classification results, with a F-measure of 0.876 in their best configuration. They also try classification for all the seven quality classes, as done in this work, using a random forest classifier with 100 trees, with a reduced set of features. The poor results (an average F-measure of 0.425) highlights the hardness of this fine-grained classification. In this paper, we address this last task in a novel way, by introducing

domain features, specially dealing with the medical domain. The results of the investigation are promising.

Recent studies specifically address the quality of medical information (in Wikipedia as well as in other resources): in [3] and [4], the authors debate if Wikipedia is a reliable learning resource for medical students, evaluating articles on respiratory topics and cardiovascular diseases. The evaluation is carried out by exploiting DISCERN<sup>14</sup>, a tool evaluating readability of articles. In [18] the authors provide novel solutions for measure the quality of medical information in Wikipedia, by adopting an unsupervised approach based on the Analytic Hierarchy Process, a multi-criteria decision making technique [22]. The work in [8] aims to provide the web surfers a numerical indication of Quality of Medical Web Sites. In particular in [8] the author proposes an index to make IQ judgment of the content and of its reliability, to give the so called “surface markers” and “trust indicator”. A similar measurement is considered in [25], where the authors present an empirical analysis that suggests the need to define genre-specific templates for quality evaluation and to develop models for an automatic genre-based classification of health information Web pages. In addition, the study shows that consumers may lack the motivation or literacy skills to evaluate the information quality of health Web pages. Clearly, this further highlights the cruciality to develop accessible automatic information quality evaluation tools and ontologies. Our work moves towards the goal, by specifically considering domain-relevant features and featuring an automatic classification task spanning over more than two classes.

## 8 Conclusions

In this work, we aimed to provide a fine grained classification mechanism for all the quality classes of the articles of the Wikipedia Medical Portal. The idea was to propose an automatic instrument for helping the reviewers to understand which articles are the less work-demanding papers to pass to next quality stage. We focused on an actionable model, namely whose features are related to the content of the articles, so that they can also directly suggest strategies for improving a given article. An important and novel aspect of our classifier, with respect to previous works, is the leveraging of features extracted from the specific, medical domain, with the help of Natural Language Processing techniques. As the results of our experiments confirm, considering specific domain-based features, like Domain Informativeness and Category, can eventually help and improve the automatic classification results. Since the results are encouraging, as future work we will evaluate other features based on the specific medical domain. Moreover, we are planning to extend our idea, to include and compare also other non medical articles (thus, extending the work to include other domains), in order to further validate our approach.

**Acknowledgments.** The research leading to these results has been partially funded by the Registro.it project My Information Bubble MIB.

---

<sup>14</sup> <http://www.discern.org.uk/>

## References

1. G. Attardi, V. Cozza, and D. Sartiano. Adapting linguistic tools for the analysis of italian medical records. *The First Italian Conference on Computational Linguistics CLiC-it 2014*, page 17, 2014.
2. G. Attardi, S. D. Rossi, and M. Simi. The TANL pipeline. *Web Services and Processing Pipelines in HLT: Tool Evaluation, LR Production and Validation (LREC:WSSP)*, 2010.
3. S. A. Azer. Is Wikipedia a reliable learning resource for medical students? Evaluating respiratory topics. *Advances in Physiology Education*, 39(1):5–14, 2015.
4. S. A. Azer, N. AlSwaidan, L. AlSwairikh, and J. AlShammari. Accuracy of cardiovascular articles on Wikipedia: Are they reliable learning resources for medical students? *BMJ Open*, bmjopen-2015-008187, 14-Mar-2015.
5. S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.
6. J. E. Blumenstock. Size matters: Word count as a measure of quality on Wikipedia. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, pages 1095–1096, New York, NY, USA, 2008. ACM.
7. O. Bodenreider and A. T. McCray. Exploring semantic groups through visual approaches. *Journal of biomedical informatics*, 36(6):414–432, 2003.
8. F. Cabitza. An information reliability index as a simple consumer-oriented indication of quality of medical web sites. In *Quality Issues in the Management of Web Information*, volume 50 of *Intelligent Systems Reference Library*, pages 159–177. Springer Berlin Heidelberg, 2013.
9. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.
10. N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6, June 2004.
11. T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.
12. G. De la Calzada and A. Dekhtyar. On measuring the quality of Wikipedia articles. In *Proceedings of the 4th Workshop on Information Credibility, WICOW '10*, pages 11–18, New York, NY, USA, 2010. ACM.
13. F. Dell'Orletta, S. Montemagni, and G. Venturi. Assessing document and sentence readability in less resourced languages and across textual genres. *International Journal of Applied Linguistics*, 31:163–193, 2014.
14. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.
15. H. He and E. Garcia. Learning from Imbalanced Data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, Sept 2009.
16. R. Hodson. Wikipedians reach out to academics. *Nature News*, Sept. 2015.
17. J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas for navy enlisted personnel. *National Technical Information Service: Research Report*, pages 8–75, 1975.
18. E. Marzini, A. Spognardi, I. Matteucci, P. Mori, M. Petrocchi, and R. Conti. Improved automatic maturity assessment of Wikipedia medical articles. In *On the Move to Meaningful Internet Systems: OTM 2014 Conferences*, volume 8841 of *Lecture Notes in Computer Science*, pages 612–622. Springer, 2014.



19. P. Nakov and T. Zesch, editors. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland, August 2014.
20. G. Pasi, G. Bordogna, and L. Jain. An introduction to quality issues in the management of web information. In G. Pasi, G. Bordogna, and L. C. Jain, editors, *Quality Issues in the Management of Web Information*, volume 50 of *Intelligent Systems Reference Library*, pages 1–3. Springer Berlin Heidelberg, 2013.
21. D. M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technologies*, 2(1):37–63, 2011.
22. T. L. Saaty. How to make a decision: The Analytic Hierarchy Process. *European Journal of Operational Research*, 48(1), 1990.
23. T. Simonite. Artificial intelligence aims to make Wikipedia friendlier and better. In *MIT Technology Review*, November 2015.
24. J. Sjöholm. Probability as readability: A new machine learning approach to readability assessment for written Swedish. Master’s thesis, Linköping University, 2012.
25. B. Stvilia, L. Mon, and Y. J. Yi. A model for online consumer health information quality. *Journal of the American Society for Information Science and Technology*, 60(9):1781–1791, 2009.
26. B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Proceedings of the 2005 International Conference on Information Quality*, pages 442–454, Cambridge, MA, 2005. MIT.
27. S. Vajjala and D. Meurers. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *International Journal of Applied Linguistics*, 165(2):194–222, 2014.
28. M. Warncke-Wang, D. Cosley, and J. Riedl. Tell me more: An actionable quality model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*, WikiSym ’13, pages 8:1–8:10, New York, NY, USA, 2013. ACM.
29. K. Wecel and W. Lewoniewski. Modelling the quality of attributes in Wikipedia infoboxes. In *Business Information Systems Workshops*, volume 228 of *Business Information Processing*, pages 308–320. Springer International Publishing, 2015.
30. K. Wu, Q. Zhu, Y. Zhao, and H. Zheng. Mining the factors affecting the quality of wikipedia articles. *Information Science and Management Engineering (ISME)*, 2010 International Conference of, 1:343–346, Aug 2010.